Contents

Description

Computational diagnosis of amino acid variants in the human exome is the first step in assessing the disruptive impacts of non-synonymous single nucleotide variants (nsSNVs) on human health and disease. MEGA-MD (**M**olecular **E**volutionary **G**enetics **A**nalysis – **M**utation **D**iagnosis) is a client-server application used to forecast the deleteriousness of nsSNVs using multiple methods and explore them in the context of the variability permitted in the long-term evolution of the affected positions.

MEGA-MD accesses a relational database (MD-DB) resident on our servers that contains pre-computed diagnoses, and associated information, for all possible mutations at all amino acid positions in the human exome. We have included three primary methods (PolyPhen-2, SIFT, and EvoD) of predicting the functional impact of amino acid variants. The first two are the most popular methods and the third significantly improves the performance for nSNVs found at ultra-conserved and at fast-evolving positions (Kumar et al., 2012). The PolyPhen-2 and SIFT diagnoses were obtained from dbNSFP. We have also included results from a multi-method consensus diagnosis, because they have been shown to be more reliable. In this case, we use the evolutionarily-balanced (see Liu and Kumar 2013) versions of PolyPhen-2 and SIFT diagnosis.

In addition to retrieving pre-computed predictions for variants in the human exome, MEGA-MD provides a facility to infer ancestral states for the position where a given amino acid mutation is found. Maximum parsimony and maximum likelihood approaches are supported by this utility which uses the 46 species reference phylogeny along with the 46 species peptide alignment for the relevant gene (obtained from the UCSC resource).

MEGA-MD is developed using the MEGA (Molecular Evolutionary Genetics Analysis) software package.

Citing MEGA-MD

Stecher G, Tamura K, Sanderford M, Peterson D, Liu L, Kumar S.

MEGA-MD: Molecular Evolutionary Genetics Analysis software for mutational diagnosis of amino acid variation.

Bioinformatics (2013) (submitted).

Disclaimer

Although the utmost care has been taken to ensure the correctness of the software, it is provided "as is," without any warranty of any kind. In no event shall the authors or their employers be considered liable for any damages, including, but not limited to, special, consequential, or other damages. The authors specifically disclaim all other warranties, expressed or implied, including, but not limited to, the determination of the suitability of this product for a specific purpose, use or application.

Copyright

Development Team

Glen Stecher[1]
- MEGA-MD desktop client
- MEGA-MDW web application

Koichiro Tamura[3]
- MEGA-MD desktop client

Maxwell Sanderford[1]
- MEGA-MDW database administration
- EvoD calculation pipeline developer

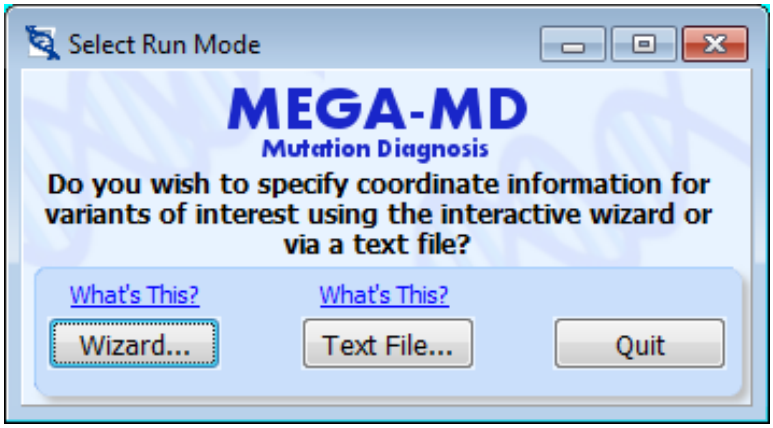
Dan Peterson[1]
- MEGA-MDW web application

Sudhir Kumar[1,2]
- Project Director

[1]Center for Evolutionary Medicine and Informatics, Biodesign Institute, Arizona State University (ASU), Tempe, AZ 85287
[2] School of Life Sciences, ASU, Tempe, AZ 85287
[3]Department of Biology, Tokyo Metropolitan University, Hachioji-shi, Tokyo 192-0397, Japan.

Modes for running MEGA-MD

When MEGA-MD is started, a prompt window is displayed that offers a choice of using an interactive Wizard for specifying variants of interest or just loading a text file with coordinate info for variants of interest formatted according to a simple format.



When using the interactive Wizard, graphical tools are used for specifying coordinate data by:

1. Searching a database for the genes which harbor variants of interest
2. Specifying amino acid positions of variants of interest using a graphical grid display of gene sequence data
3. Selecting amino acid mutations from a list

When loading coordinate information via a text file, the information in the file is validated and then predictions are automatically retrieved from the MEGA-MDW server and displayed in the Mutation Explorer window.

Interactive Wizard

MEGA-MD provides an interactive system for specifying the coordinate information for variants of interest when few variants are to be explored (see here if exploring many variants). When using the interactive system, the Mutation Explorer window is displayed with the Gene Search tab selected. In the search box on this tab, you can enter a gene name, gene product description, RefSeq mRNA id or peptide id and then press enter or click the search button. A request will then be sent to the MEGA-MDW web server for a list of genes that match the given query. All matches are then displayed in a table view below the search box. This table view displays the gene name, peptide id, and gene product for all results returned by the MEGA-MDW web server.

To specify the amino acid position and mutant amino acid(s), click the *Diagnose Variant* link in the same row that lists the gene which harbors the variant of interest. After clicking this link, the UCSC 46-species peptide alignment for the specified gene will be retrieved from the MEGA-MDW web server and displayed in the Sequence Data Explorer window. In this window, you first specify the amino acid position for the variant of interest by clicking at that location in the alignment grid. When doing so, that column in the alignment grid will be highlighted and the position will be indicated on the *Diagnose Variant* button. After the position is specified, the mutant amino acid is specified by clicking the *Diagnose Variant* button which is located on the toolbar. A drop down list is then displayed from which you can select a single mutant amino acid or all possibilities.

Once the amino acid position in the alignment and the mutant amino acid have been designated, predictions and related data for that variant will be retrieved from the MEGA-MDW web server and displayed in the Mutation Explorer window.

Load a text file with coordinate info

When retrieving predictions for a significant number of variants, the interactive Wizard system is cumbersome and so MEGA-MD provides a second way to specify coordinate

information for variants of interest. Instead of using the interactive Wizard, you can load a text file with coordinate information as follows:

Create a text file with coordinate information for all nsSNVs to be explored following the format below:

```
NP_000758 99 E
NP_000761 264 M
NP_000762 144 C
NP_000762 335 W
NP_000773 374 T
NP_000838 71 L
NP_000886 131 H
NP_000887 271 T
```

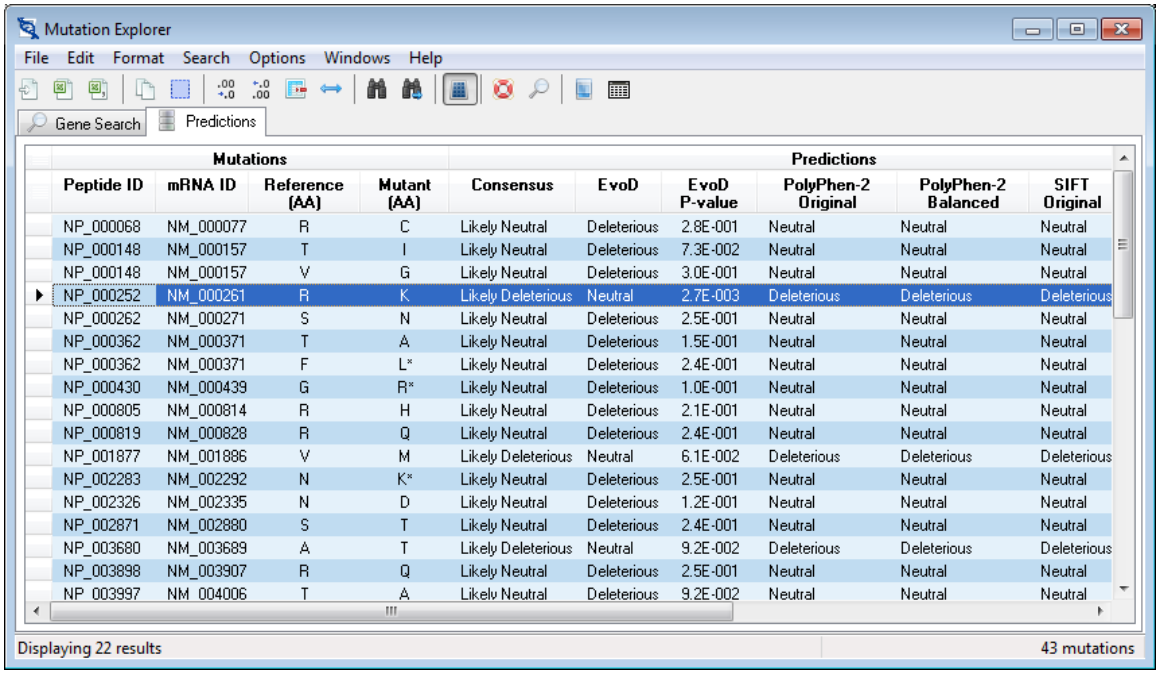
Each line contains coordinate information for one nsSNV and each value is separated by white space (*i.e.* spaces or tabs).

In the Mutation Explorer window, select *File->Import Query Data From File* (or click the upload data button) and browse for the newly created text file. MEGA-MD will first validate the format of the coordinate information file and then request prediction information for all specified nsSNVs from the MEGA-MDW web server. As data is retrieved, the Mutation Explorer window is updated.

The MEGA-MD application has no limit on the number of entries that can be included in the coordinate information file. However, depending on your internet connection speed and the current load on the MEGA-MDW server, retrieval of many predictions may take some time (anything less than 5,000 should not be problematic). For situations where MEGA-MD does not perform optimally due to high numbers of nsSNVs, the MEGA-MDW can be used directly (www.mypeg.info/MEGA-MDW). The same text file can be uploaded to the MEGA-MD server which will process the file and send you an email for retrieving prediction data once the processing is complete.

Mutation Explorer

The *Mutation Explorer* window displays predictions and data associated with the nsSNVs being explored and provides functionality for text searching, sorting, importing, exporting, formatting, and gene search. This window displays two main views, each located on a separate tab:



Mutations				Predictions					
Peptide ID	mRNA ID	Reference (AA)	Mutant (AA)	Consensus	EvoD	EvoD P-value	PolyPhen-2 Original	PolyPhen-2 Balanced	SIFT Original
NP_000068	NM_000077	R	C	Likely Neutral	Deleterious	2.8E-001	Neutral	Neutral	Neutral
NP_000148	NM_000157	T	I	Likely Neutral	Deleterious	7.3E-002	Neutral	Neutral	Neutral
NP_000148	NM_000157	V	G	Likely Neutral	Deleterious	3.0E-001	Neutral	Neutral	Neutral
NP_000252	NM_000261	R	K	Likely Deleterious	Neutral	2.7E-003	Deleterious	Deleterious	Deleterious
NP_000262	NM_000271	S	N	Likely Neutral	Deleterious	2.5E-001	Neutral	Neutral	Neutral
NP_000362	NM_000371	T	A	Likely Neutral	Deleterious	1.5E-001	Neutral	Neutral	Neutral
NP_000362	NM_000371	F	L*	Likely Neutral	Deleterious	2.4E-001	Neutral	Neutral	Neutral
NP_000430	NM_000439	G	R*	Likely Neutral	Deleterious	1.0E-001	Neutral	Neutral	Neutral
NP_000805	NM_000814	R	H	Likely Neutral	Deleterious	2.1E-001	Neutral	Neutral	Neutral
NP_000819	NM_000828	R	Q	Likely Neutral	Deleterious	2.4E-001	Neutral	Neutral	Neutral
NP_001877	NM_001886	V	M	Likely Deleterious	Neutral	6.1E-002	Deleterious	Deleterious	Deleterious
NP_002283	NM_002292	N	K*	Likely Neutral	Deleterious	2.5E-001	Neutral	Neutral	Neutral
NP_002326	NM_002335	N	D	Likely Neutral	Deleterious	1.2E-001	Neutral	Neutral	Neutral
NP_002871	NM_002880	S	T	Likely Neutral	Deleterious	2.4E-001	Neutral	Neutral	Neutral
NP_003680	NM_003689	A	T	Likely Deleterious	Neutral	9.2E-002	Deleterious	Deleterious	Deleterious
NP_003898	NM_003907	R	Q	Likely Neutral	Deleterious	2.5E-001	Neutral	Neutral	Neutral
NP_003997	NM_004006	T	A	Likely Neutral	Deleterious	9.2E-002	Neutral	Neutral	Neutral

The actions provided by the *Mutation Explorer* are divided into several categories and are accessed using the main menu bar or the main tool bar:

File

- Import Query Data From File – load coordinate information form a text file
- Search for a Gene – access the gene search page
- Export Table to Excel File – save all prediction data to an MS Excel file
- Export Table to CSV File – save all prediction data to a Comma-Separated-Values text file
- Exit – Close the application

Edit

- Copy – copy selected values to the system clip-board
- Select All – select all values in the table
- Clear Table – clear all data from the table

Format

- Increase Precision – increase the precision of all numeric values in the table (and also in the *Mutation Detail View* window)
- Decrease Precision - decrease the precision of all numeric values in the table (and also in the *Mutation Detail View* window)
- Resize Columns to Best-fit – resizes all columns in the table to achieve the best fit and optimize the view. Useful when hiding/showing columns and column widths change sub-optimally. ***note: if there are many records in the table (more than several thousand), this operation may take a few moments or more, during which time the window will be unresponsive.

Search

- Find… - text search for values in the table
- Find Next – find the next value matching the search query (search goes to the right and then down to the next row)

Options

- Keep detail view on top – toggle this action on/off to keep the *Mutation Detail View* window from staying in front of other MEGA-MD windows (on by default).
- Show Toolbar – toggle on/off the display of the toolbar (on by default)
- Toggle Auto Column Width – when off (default) a horizontal scroll bar is used to view columns that don't fit in the window. When off, the horizontal scroll bar is removed and all columns are squeezed into view.

Windows

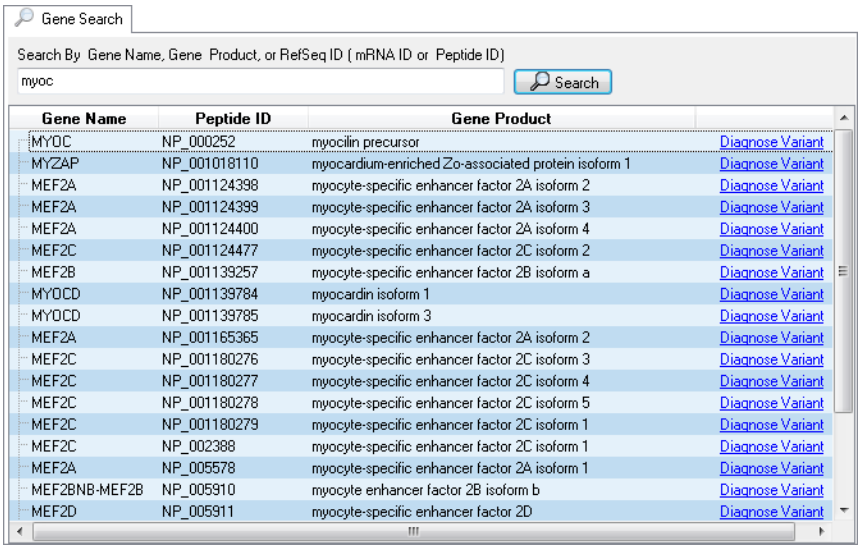
- Detail View Form – show the *Mutation Detail View* window
- Search for a Gene – jump to the *Gene Search* tab in the *Mutation Explorer* window
- Sequence Data Explorer – show the *Sequence Data Explorer* window

Help

- Contents – Display this help document
- About – show the *About MEGA-MD* window

Gene Search tab

The *Gene Search* tab facilitates searching for genes by name, keyword (based on gene product), or by RefSeq identifiers (mRNA ID or Protein ID). Search results (limited to 1000) are displayed in a list view with cursory information and a link for retrieving the 46-species reference protein sequence alignment from the MEGA-MDW server. When a sequence alignment is retrieved it is displayed in the Sequence Data Explorer which can be used to specify the amino acid site and mutant allele for a nsSNV of interest.



Gene Name	Peptide ID	Gene Product	
MYOC	NP_000252	myocilin precursor	Diagnose Variant
MYZAP	NP_001018110	myocardium-enriched Zo-associated protein isoform 1	Diagnose Variant
MEF2A	NP_001124398	myocyte-specific enhancer factor 2A isoform 2	Diagnose Variant
MEF2A	NP_001124399	myocyte-specific enhancer factor 2A isoform 3	Diagnose Variant
MEF2A	NP_001124400	myocyte-specific enhancer factor 2A isoform 4	Diagnose Variant
MEF2C	NP_001124477	myocyte-specific enhancer factor 2C isoform 2	Diagnose Variant
MEF2B	NP_001139257	myocyte-specific enhancer factor 2B isoform a	Diagnose Variant
MYOCD	NP_001139784	myocardin isoform 1	Diagnose Variant
MYOCD	NP_001139785	myocardin isoform 3	Diagnose Variant
MEF2A	NP_001165365	myocyte-specific enhancer factor 2A isoform 2	Diagnose Variant
MEF2C	NP_001180276	myocyte-specific enhancer factor 2C isoform 3	Diagnose Variant
MEF2C	NP_001180277	myocyte-specific enhancer factor 2C isoform 4	Diagnose Variant
MEF2C	NP_001180278	myocyte-specific enhancer factor 2C isoform 5	Diagnose Variant
MEF2C	NP_001180279	myocyte-specific enhancer factor 2C isoform 1	Diagnose Variant
MEF2C	NP_002388	myocyte-specific enhancer factor 2C isoform 1	Diagnose Variant
MEF2A	NP_005578	myocyte-specific enhancer factor 2A isoform 1	Diagnose Variant
MEF2BNB-MEF2B	NP_005910	myocyte enhancer factor 2B isoform b	Diagnose Variant
MEF2D	NP_005911	myocyte-specific enhancer factor 2D	Diagnose Variant

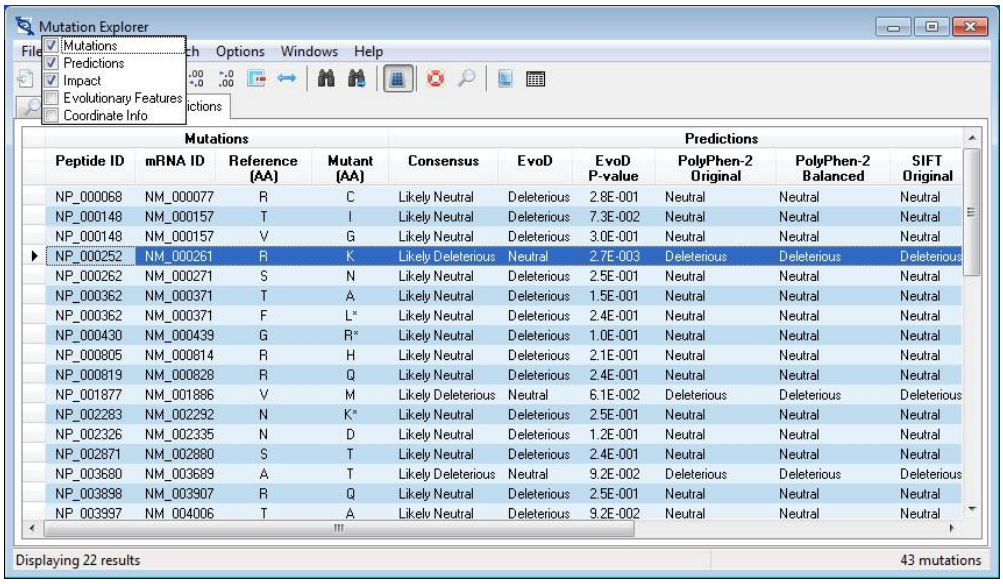
Predictions

The *Predictions* tab displays all prediction data retrieved from the MEGA-MDW server in a list view. Complete information for the currently active record is displayed in the Mutation Detail View. Columns of data are banded together into categories:

- Mutations – identifiers as well as mutant and reference alleles are given here. Note – mutant amino acids that are appended with an asterisk (*) have multiple rows returned by the MEGA-MD server, each row indicating a mutation at the nucleotide level (look to the Coordinate Info band to see nucleotide change).
- Predictions – consensus, EvoD, PolyPhen-2, and SIFT predictions are given here. Where both the original and balanced predictions are given for PolyPhen-2 and SIFT (balanced predictions are described in Liu and Kumar 2013).
- Impact – the impact scores for EvoD, PolyPhen-2, and SIFT predictions are provided along with the Grantham distance and Blosum62 value.
- Evolutionary Features (hidden by default) – substitution rate, position time span, and mutation time span are displayed (see below for a description of how to display this band).
- Coordinate Info (hidden by default) – additional coordinate information is shown here, including chromosome, strand, nucleotide position, amino acid position, wild nucleotide, and mutant nucleotide (see below for a description of how to display this band).

To toggle on/off the display of a given band, click on the indicator button which is located to the far left in the band headers row. A popup menu will appear from which bands can be selected/deselected. Often times when changing the display of bands, column widths will change in undesirable ways. To remedy this, you can execute the *Best-fit Columns* action by clicking *Format->Resize columns to best-fit* or clicking the toolbar button. Alternatively, columns widths can be adjusted by dragging their header edges.

The toolbar and main menu provide access to several actions for importing/exporting data, formatting the view, sorting, text search, and setting view options.



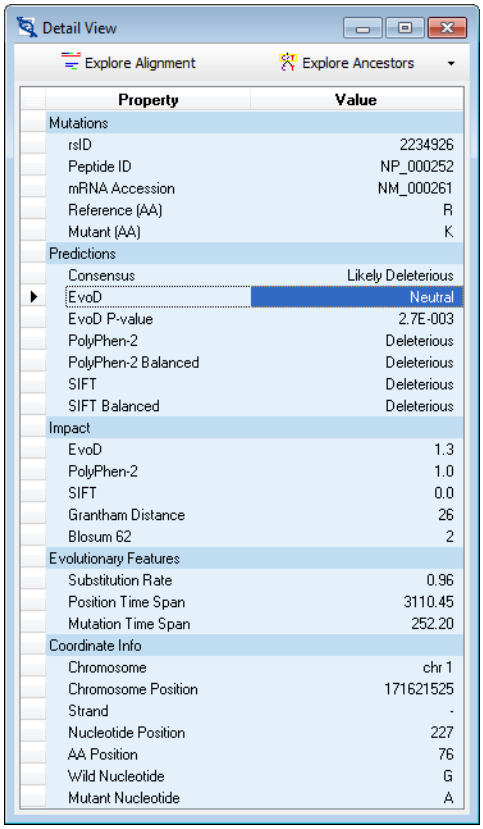
Mutations				Predictions					
Peptide ID	mRNA ID	Reference (AA)	Mutant (AA)	Consensus	EvoD	EvoD P-value	PolyPhen-2 Original	PolyPhen-2 Balanced	SIFT Original
NP_000068	NM_000077	R	C	Likely Neutral	Deleterious	2.8E-001	Neutral	Neutral	Neutral
NP_000148	NM_000157	T	I	Likely Neutral	Deleterious	7.3E-002	Neutral	Neutral	Neutral
NP_000148	NM_000157	V	G	Likely Neutral	Deleterious	3.0E-001	Neutral	Neutral	Neutral
NP_000252	NM_000261	R	K	Likely Deleterious	Neutral	2.7E-003	Deleterious	Deleterious	Deleterious
NP_000262	NM_000271	S	N	Likely Neutral	Deleterious	2.5E-001	Neutral	Neutral	Neutral
NP_000362	NM_000371	T	A	Likely Neutral	Deleterious	1.5E-001	Neutral	Neutral	Neutral
NP_000362	NM_000371	F	L*	Likely Neutral	Deleterious	2.4E-001	Neutral	Neutral	Neutral
NP_000430	NM_000439	G	R*	Likely Neutral	Deleterious	1.0E-001	Neutral	Neutral	Neutral
NP_000805	NM_000814	R	H	Likely Neutral	Deleterious	2.1E-001	Neutral	Neutral	Neutral
NP_000819	NM_000828	R	Q	Likely Neutral	Deleterious	2.4E-001	Neutral	Neutral	Neutral
NP_001877	NM_001886	V	M	Likely Deleterious	Neutral	6.1E-002	Deleterious	Deleterious	Deleterious
NP_002283	NM_002292	N	K*	Likely Neutral	Deleterious	2.5E-001	Neutral	Neutral	Neutral
NP_002326	NM_002335	N	D	Likely Neutral	Deleterious	1.2E-001	Neutral	Neutral	Neutral
NP_002871	NM_002880	S	T	Likely Neutral	Deleterious	2.4E-001	Neutral	Neutral	Neutral
NP_003680	NM_003689	A	T	Likely Deleterious	Neutral	9.2E-002	Deleterious	Deleterious	Deleterious
NP_003898	NM_003907	R	Q	Likely Neutral	Deleterious	2.5E-001	Neutral	Neutral	Neutral
NP_003997	NM_004006	T	A	Likely Neutral	Deleterious	9.2E-002	Neutral	Neutral	Neutral

Mutation Detail View

The Mutation Detail View window displays all available information for the currently active record (selected in the Mutation Explorer window). Additionally, this window provides access to the 46-species reference alignment for the given gene as well as the ability to infer ancestral alleles using the Maximum Likelihood (ML) or Maximum Parsimony (MP) methods.

When the *Explore Alignment* button is clicked, MEGA-MD will retrieve the 46-species reference alignment from the MEGA-MDW server and display it in the Sequence Data Explorer, from where it can be exported or further exploration can be done.

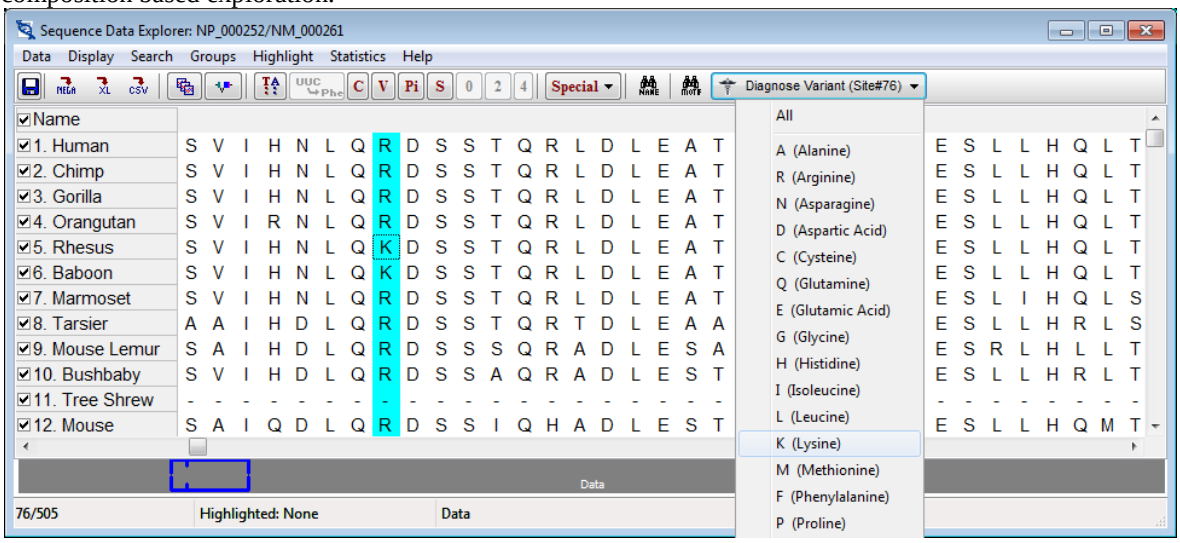
When the *Explore Ancestors* button is clicked, the choice of ML and MP methods are presented. If the ML approach is selected, the Analysis Preferences Dialog is displayed from which the analysis can be launched with custom settings (e.g. substitution model, distribution of rates, etc…). If the MP approach is selected, the analysis is launched immediately as not custom settings are available for this method. When the analysis is completed, the reference topology will be displayed in the Tree Explorer along with inferred ancestral alleles for the amino acid site designated earlier.



Sequence Data Explorer

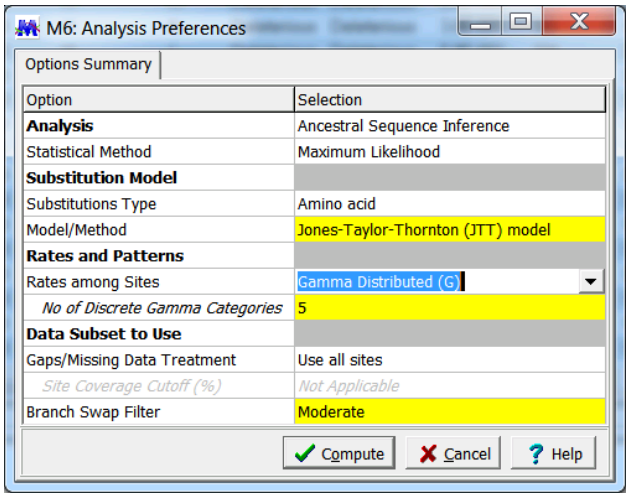
The *Sequence Data Explorer* is used to display the 46-species alignment for a given gene and provides a graphical interface for specifying amino acid position and mutant allele for nsSNVs of interest. With an alignment activated, the amino acid position is specified by selecting the site of interest (which will be highlighted). With the site of interest selected, the mutant allele (or all alleles) can be specified from the *Diagnose Variant* drop down list. When an allele is selected from the list, MEGA-MD will query the MEGA-MDW server and append the returned predictions and related data to the Mutation Explorer *Predictions* tab.

The *Sequence Data Explorer* window also provides much other functionality such as alignment export and composition based exploration.



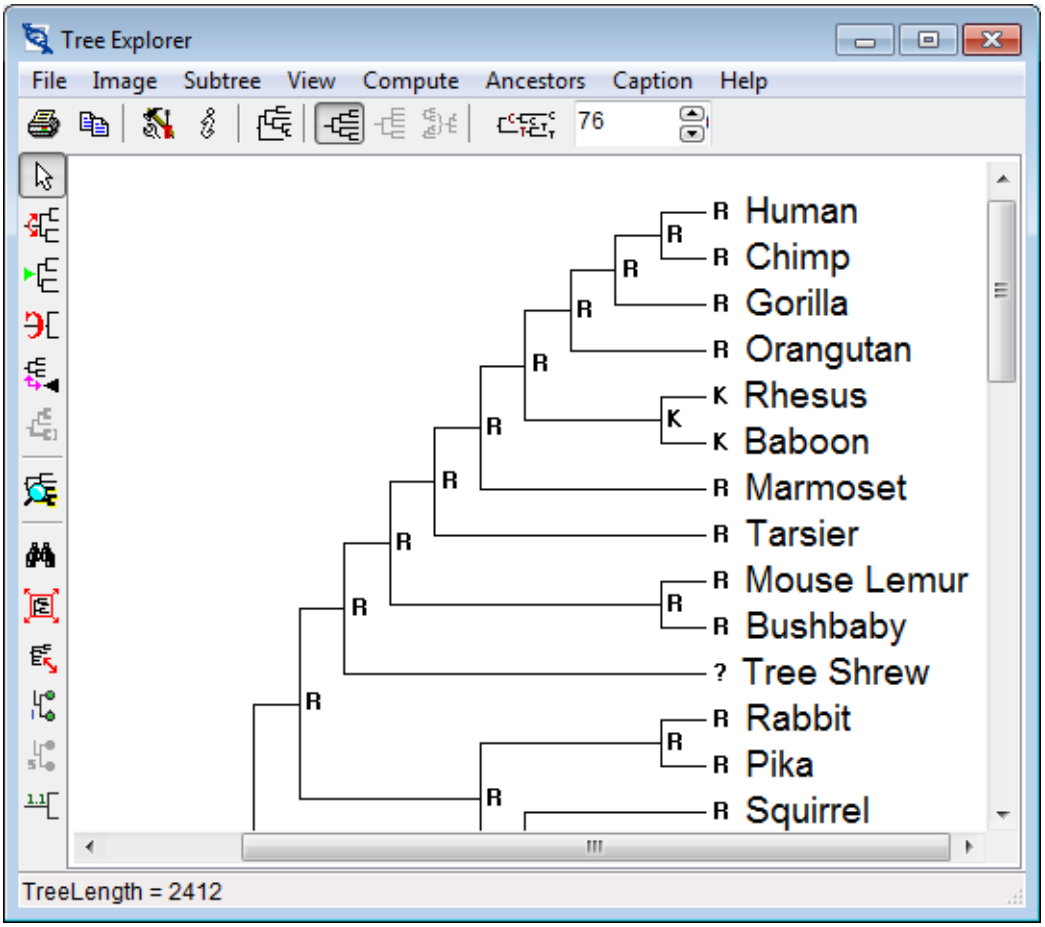
Analysis Preferences Dialog

The *Analysis Preferences Dialog* is used for specifying the substitution model to use as well as the distribution of rates for ML based ancestral sequence inference.



Tree Explorer

The Tree Explorer displays the results of the ancestral sequence inference analyses. When an ancestral sequence inference analysis is complete, the 46-species reference phylogeny is displayed in the Tree Explorer with the inferred ancestral alleles for the nsSNV amino acid site shown. The Tree Explorer provides many capabilities which are described in detail in the MEGA user manual which can be obtained from www.megasoftware.net/manual.pdf .



Input Data Overview

In order to retrieve predictions for a given nsSNV, MEGA-MD requires three pieces of information:

1. RefSeq protein id (e.g. NP_000082)
2. amino acid position (e.g. 43)
3. mutant allele (e.g. R)

There are two ways to provide this coordinate information to MEGA-MD

Upload a text file

Use the interactive Wizard (via Gene Search and integrated Sequence Data Explorer)

Upload a text file with the coordinate information for all nsSNVs of interest

Create a text file with coordinate information for all nsSNVs to be explored following the format below:

NP_000758 99 E
NP_000761 264 M
NP_000762 144 C
NP_000762 335 W
NP_000773 374 T
NP_000838 71 L
NP_000886 131 H
NP_000887 271 T

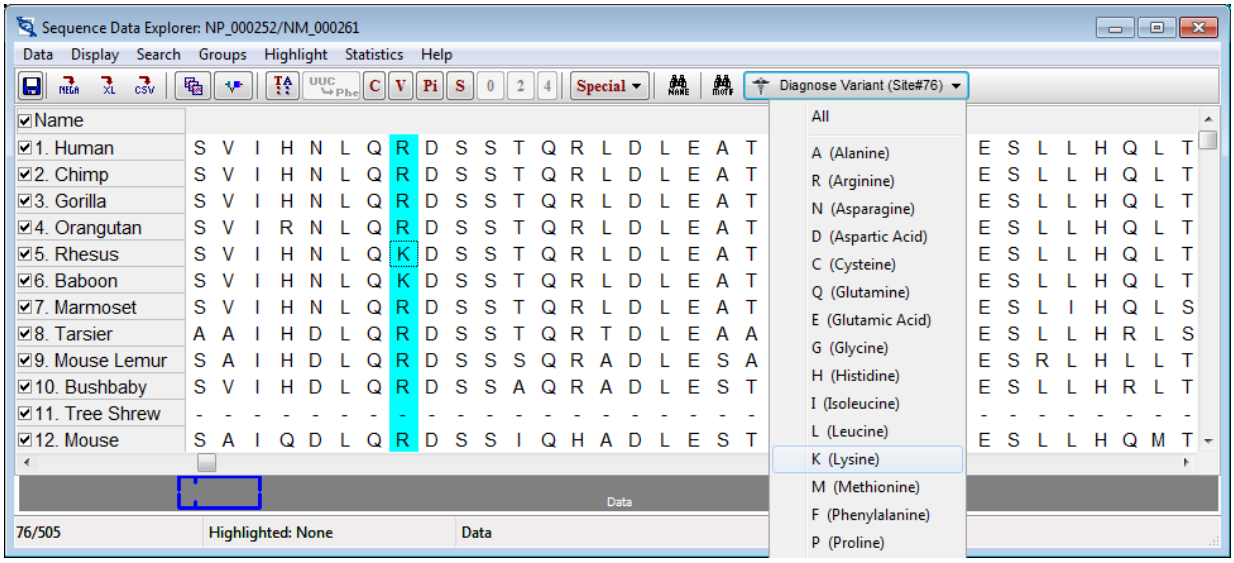
Each line contains coordinate information for one nsSNV and each value is separated by white space (*i.e.* spaces or tabs).

In the Mutation Explorer window, select *File->Import Query Data From File* (or click the upload data button) and browse for the newly created text file. MEGA-MD will first validate the format of the coordinate information file and then request prediction information for all specified nsSNVs from the MEGA-MDW web server. As data is retrieved, the Mutation Explorer window is updated.

The MEGA-MD application has no limit on the number of entries that can be included in the coordinate information file. However, depending on your internet connection speed and the current load on the MEGA-MDW server, retrieval of many predictions may take some time (anything less than 5,000 should not be problematic). For situations where MEGA-MD does not perform optimally due to high numbers of nsSNVs, the MEGA-MDW can be used directly (www.mypeg.info/MEGA-MDW). The same text file can be uploaded to the MEGA-MD server which will process the file and send you an email for retrieving prediction data once the processing is complete.

Specify the coordinate information using the integrated Sequence Data Explorer

If a 46-species sequence alignment has been retrieved (see Gene Search) for a given gene, the Sequence Data Explorer window can be used to first navigate to the amino acid site of interest and then specify a mutant allele.



References

dbNSFP

Liu X, Jian X, Boerwinkle E.
dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions.
Hum Mutat. (2011);32(8):894-9. doi: 10.1002/humu.21517.

MEGA-MD

Kumar S, Sanderford M, Gray VE, Ye J, Liu Li.
Evolutionary diagnosis method for variants in personal exomes.
Nature Methods (2012) Sep;9(9):855-6. doi:10.1038/nmeth.2147.

Liu L. and Kumar S.
Evolutionary balancing is critical for correctly predicting amino acid variants with functional impact. *Molecular Biology and Evolution* (2013), 30, 1252-1257.

MEGA (Molecular Evolutionary Genetics Analysis)

Tamura K, Stecher G, Peterson D, Filipski A, and Kumar S.
MEGA6: Molecular Evolutionary Genetics Analysis version 6.0
Molecular Biology and Evolution(2013), 30, 2725-2729.

PolyPhen2

Adzhubei IA, Schmidt S, Peshkin L, Ramensky V, Gerasimova A, Bork P, Kondrashov A, Sunyaev S.
A method and server for predicting damaging missense mutations. *Nature Methods* (2010) 7: 248-249.

SIFT

Pauline C. Ng and Steven Henikoff
SIFT: predicting amino acid changes that affect protein function. *Nucl. Acids Res.* (2003) 31(13): 3812-3814 doi:10.1093/nar/gkg509

UCSC Genome Browser Database

Fujita, P.A., et al.
The UCSC Genome Browser database: update 2011
Nucleic Acids Research (2011), 39, D876-D882.